Comparison of Biological Significance of Biclusters of SIMBIC and SIMBIC+ Biclustering Models

J.Bagyamani[1], K.Thangavel[2] and R.Rathipriya[2]
[1] Government Arts College/Computer Science, Dharmapuri, India
[2] Periyar University/Computer Science, Salem, India
Email: [1] bagya.gac@gmail.com,[2] drktvelu@yahoo.com, [3] rathipriyar@yahoo.co.in

***Abstract*— Query driven Biclustering Model refers to the problem of extracting biclusters based on a query gene or query condition. The extracted biclusters consist of a set of genes and a subset of conditions that are similar to the query gene or query condition and it includes the query input also. Two approaches applied for biclustering problems are *top-down* and *bottom-up*, based on how they tackle the problems. Top-down techniques [3, 4] start with the entire gene expression matrix and iteratively partition it into smaller sub-matrices. On the other hand, bottom-up approach starts with a randomly chosen set of biclusters that are iteratively modified, usually enlarged, until no local improvement is possible. In this paper, the biological significance of biclusters extracted using two query driven models viz SIMBIC and SIMBIC+ are compared.This paper is organized as follows. Section 2 analyzes the popular MSB algorithm and section 3 introduces an improved version of MSB namely SIMBIC model and the enhanced model of SIMBIC namely SIMBIC+ is presented in section 4. The experimental analysis and the biological significance are illustrated in section 5.**

***Index Terms* - Data Mining, Gene Expression Data, Biclustering, Average Correlation Value, Biological Significance, Gene Ontology**

I. Introduction

Existing biclustering models for microarray data analysis often do not answer the specific questions of interest to a biologist. This lack of sharpness has prevented them from surpassing a rather vague exploratory role. Often, biologists have at hand a specific gene or set of genes (seed genes) which they know or expect to be related to some common biological pathway or function. In particular, this problem formulation necessitates various questions or queries, such as 'which genes involved in a specific protein complex are co-expressed?

Thus the biclustering problem is mathematically defined as follows. Let E (G, C) be the expression matrix of size m x n. Let g_i be the query gene of interest. The biclustering problem is to find $G' \subset G$ and $C' \subset C$ such that the bicluster B = E (G', C') forms significant pattern [5].

II. MSB Algorithm

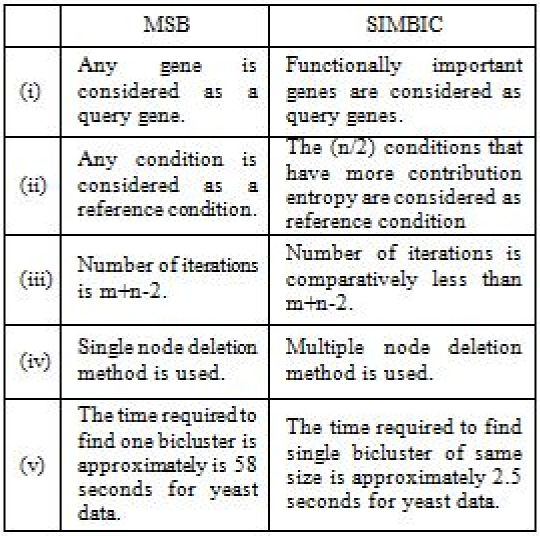
Liu X. and Wang L. [4] developed a query driven biclustering model namely Maximum Similarity Bicluster (MSB), to find an optimal bicluster with the maximum similarity score. A query gene or set of genes is given as input. The model constructs a similarity matrix S(G, C) based on similarity between the query gene and the other remaining genes. A gene or condition with low similarity is eliminated in each cycle until the similarity matrix reduces to a single element. Then a bicluster with maximum similarity is extracted. This algorithm could extract constant and additive biclusters. Biclusters of MSB were extracted using BicAT-plus to compare different biclustering methods based on biological merits.

III. Simbic Model

The MSB algorithm is improved by introducing t-test based gene selection, contribution to the entropy based condition selection and multiple node deletion techniques. Similarity score between genes and similarity score for a bicluster are defined as in MSB [4].

Multiple genes or multiple conditions with very low similarity are removed in every cycle until the similarity matrix reduces to a single element. Then a bicluster with maximum similarity is extracted. The comparison of MSB and SIMBIC biclustering models is tabulated in Table I.

Table I. Comparison of MSB with SIMBIC Model



	MSB	SIMBIC
(i)	Any gene is considered as a query gene.	Functionally important genes are considered as query genes.
(ii)	Any condition is considered as a reference condition.	The (n/2) conditions that have more contribution entropy are considered as reference condition
(iii)	Number of iterations is m+n-2.	Number of iterations is comparatively less than m+n-2.
(iv)	Single node deletion method is used.	Multiple node deletion method is used.
(v)	The time required to find one bicluster is approximately is 58 seconds for yeast data.	The time required to find single bicluster of same size is approximately 2.5 seconds for yeast data.

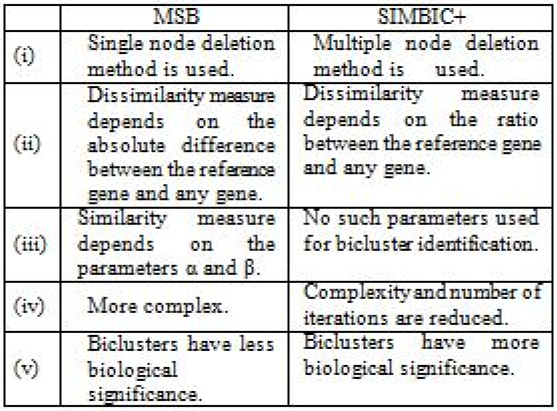
IV. Simbic+ Model

SIMBIC+ biclustering model is an enhancement of SIMBIC model in which the similarity between two genes is defined based on the ratio between the genes [2]. This ratio-based similarity measure extracts scaling pattern biclusters rather than SIMBIC which extracts constant and additive

DOI: 01.IJIT.3.1.1015

biclusters. The comparison of SIMBIC and SIMBIC+ biclustering models is depicted in Table II.

TABLE II. COMPARISON OF SIMBIC+ MODEL WITH MSB



	MSB	SIMBIC+
(i)	Single node deletion method is used.	Multiple node deletion method is used.
(ii)	Dissimilarity measure depends on the absolute difference between the reference gene and any gene.	Dissimilarity measure depends on the ratio between the reference gene and any gene.
(iii)	Similarity measure depends on the parameters α and β.	No such parameters used for bicluster identification.
(iv)	More complex.	Complexity and number of iterations are reduced.
(v)	Biclusters have less biological significance.	Biclusters have more biological significance.

V. EXPERIMENTAL ANALYSIS

In this section, the performance of SIMBIC and SIMBIC+ bicluster models are evaluated. Since microarray data has large number of features (genes), the majority of which are not relevant to the description of the problem, it could potentially degrade the gene expression analysis by masking the contribution of the relevant features. Hence after applying t-test based gene selection and contribution to the entropy based condition selection [6], experiments were conducted on benchmark Yeast Saccharomyces cerevisae gene expression dataset. The yeast cell cycle expression dataset is a time series data that contains 2,884 genes and 17 conditions. Analysis of this dataset also helps in antifungal drug discovery.

A. Bicluster Evaluation Measures

Two types of measures namely quantitative measures and qualitative measures are used to evaluate biclusters. The quantitative measures are used to quantify a bicluster in terms of size or volume of bicluster. The qualitative measures are used to determine the quality of extracted biclusters in terms of statistical measures namely variance, Mean Squared Reside (MSR) and Average Correlation Value (ACV) [7].

Statistical measures evaluate a bicluster theoretically, but the biological significance proves the real quality of the extracted bicluster. The Gene Ontology (GO) tool renders the biological significance in terms of function of the genes in the bicluster. To determine the statistical significance of the association of a particular GO term with a group of genes in the list, GO tool estimates the p-value.Though a number of tools are available to find the Gene Ontology GOTermFinder Tool has been used in this paper to discover the biological significance of the genes in the bicluster. Apart from p-value, Gene Ontology includes Biological Process (BP), Molecular Function (MF) and Cellular Component (CC) of the genes in the bicluster. Biological process refers to a biological objective to which the gene or gene product contributes. Molecular function is deûned as the biochemical activity of a gene product. Cellular component refers to the place in the cell where a gene product is active. Lower p-value confirms that the genes in the extracted bicluster are biologically significant. The significance of p-value is measured at 0.1, 0.05 and 0.01 levels.SIMBIC algorithm extracts maximum similarity bicluster corresponding to the query gene. It is possible to extract both constant and additive biclusters using this algorithm. It is observed that the number of iterations in order to extract maximum similarity bicluster is reduced considerably [1]. Thus for the query gene 288 (YBR198C) and for the default parameter setting α = 0.4, β = 0.5 and γ = 1.2, MSB extracts a constant bicluster of size 225 (15 x 15). The genes in the bicluster are YAL041W, YBR123C, YBR198C, YDL045C, YGR083C, YHL023C, YHR201C, YJR129C, YLL043W, YLR215C, YLR317W, YLR324W, YLR425W, YMR176W, YPL002C and the conditions are 2, 3, 4, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16 and 17. The Parallel Coordinate (PC) plot of constant bicluster of MSB using BicAT is provided in Fig 1.

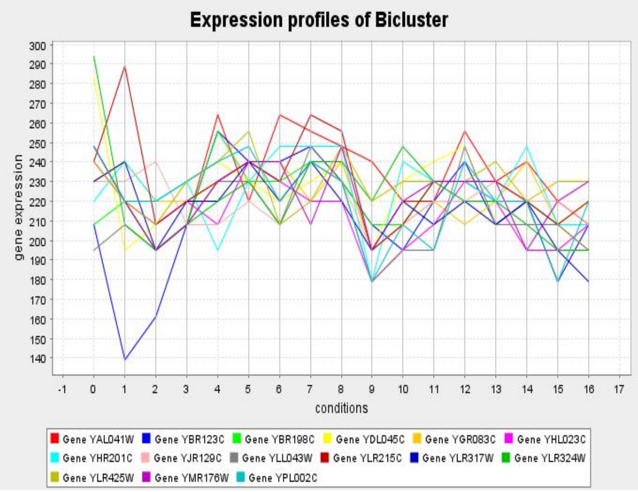


Fig. 1. PC plot of constant bicluster with query gene YBR198C using BicAT

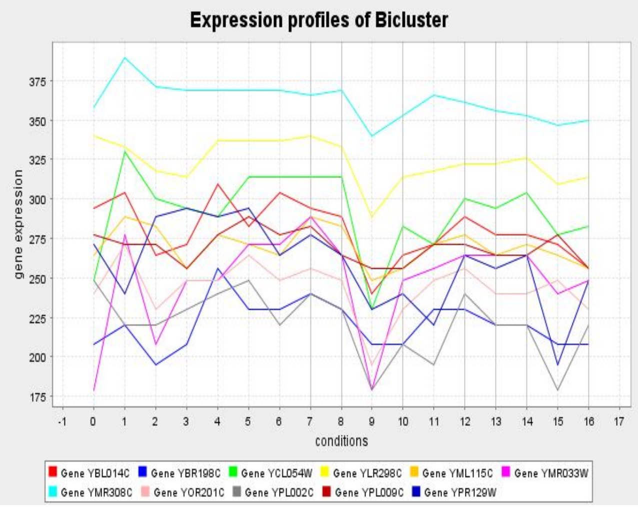


Fig. 2. PC plot of additive bicluster with query gene YBR198C using BicAT

For the same reference gene YBR198C and reference condition 14 the size of additive bicluster using BicAT is 77

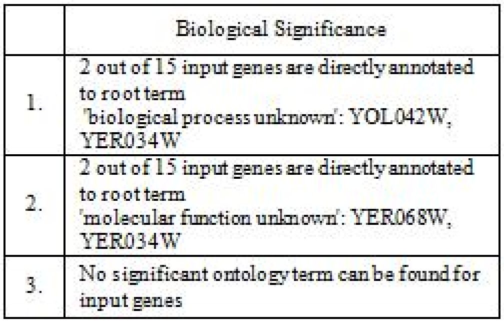
DOI: 01.IJIT.3.1.1015

(11 x 7). The PC plot of this additive bicluster is provided in Fig 2. The genes in the bicluster are YBL014C, YBR198C, YCL054W, YLR298C, YML115C, YMR033W, YMR308C, YOR201C, YPL002C, YPL009C, YPR129W and the conditions are 8, 9, 11, 13, 14, 15 and 17.

B. Comparison of Biological Significance

The biological significance of constant bicluster for the reference gene YBR198C, identified by Liu X. et al. [4] is tabulated in Table III. The biological process of the additive bicluster extracted using BicAT for the same reference gene is given in Table IV. The molecular function and cellular component of the same additive bicluster is presented in Table V and Table VI respectively.

Table III. Biological Significance of Constant Bicluster of MSB

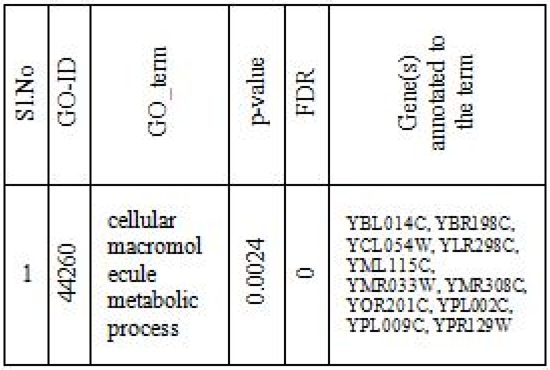


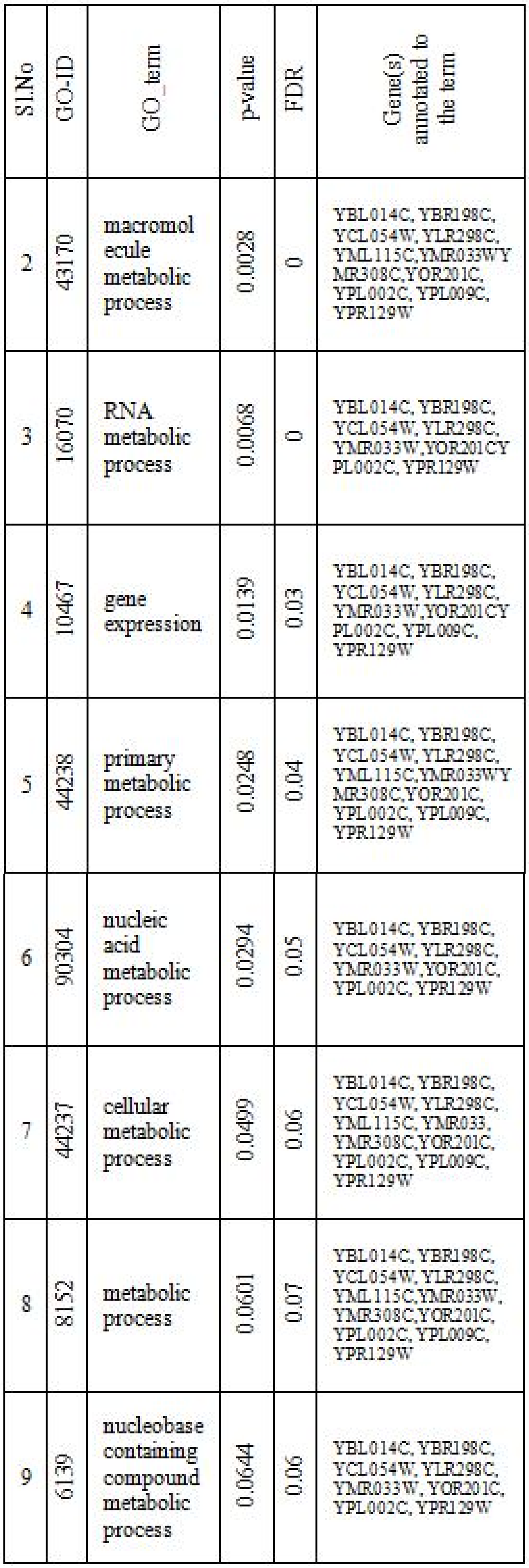
	Biological Significance
1.	2 out of 15 input genes are directly annotated to root term 'biological process unknown': YOL042W, YER034W
2.	2 out of 15 input genes are directly annotated to root term 'molecular function unknown': YER068W, YER034W
3.	No significant ontology term can be found for input genes

Thus it is observed from Table III that for the constant bicluster there is no biological significance. Since SIMBIC and MSB extract identical biclusters they have identical Gene Ontology. It is evident from Table IV that the additive bicluster has 9 significant ontologies related to the biological process. It is observed from Table V that the extracted additive bicluster has 6 significant molecular functions. Also Table VI shows that there is only one significant gene ontology related to cellular component. Thus, it is evident that additive biclusters and scaling pattern biclusters have more biological significance than constant biclusters.

Comparison of GO enrichment of biclusters of Yeast dataset obtained using SIMBIC+ and MSB is tabulated in Table VII. Since ratio based similarity is used in SIMBIC+, it is possible to extract scaling pattern biclusters. It is observed from Table VII that the biclusters extracted using SIMBIC+ have more biological significance than the biclusters of MSB.

Table IV. Biological Process of Additive Bicluster





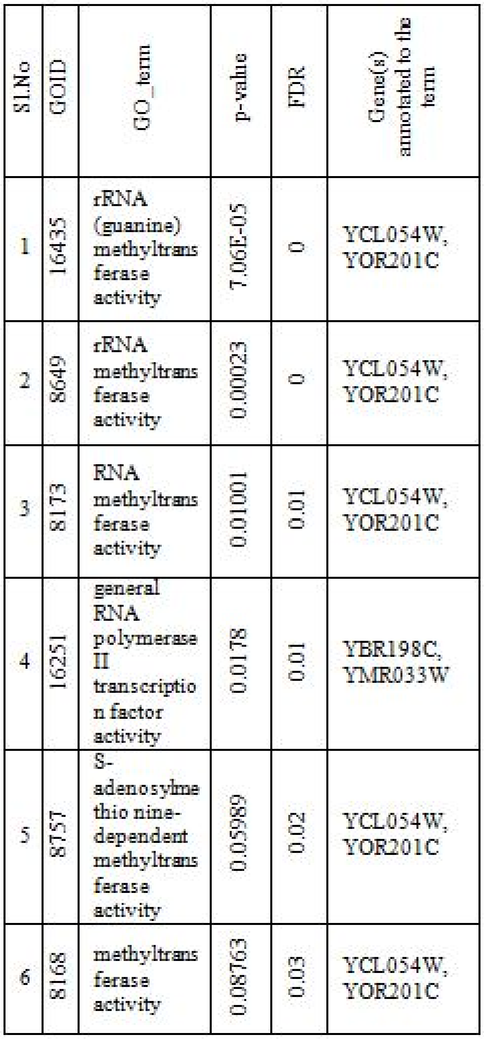
Sl.No	GO-ID	GO_term	p-value	FDR	Gene(s) annotated to the term
1	44260	cellular macromolecule metabolic process	0.0024	0	YBL014C, YBR198C, YCL054W, YLR298C, YML115C, YMR033W, YMR308C, YOR201C, YPL002C, YPL009C, YPR129W
2	43170	macromolecule metabolic process	0.0028	0	YBL014C, YBR198C, YCL054W, YLR298C, YML115C,YMR033WYMR308C,YOR201C, YPL002C, YPL009C, YPR129W
3	16070	RNA metabolic process	0.0068	0	YBL014C, YBR198C, YCL054W, YLR298C, YMR033W,YOR201CYPL002C, YPR129W
4	10467	gene expression	0.0139	0.03	YBL014C, YBR198C, YCL054W, YLR298C, YMR033W,YOR201CYPL002C, YPL009C, YPR129W
5	44238	primary metabolic process	0.0248	0.04	YBL014C, YBR198C, YCL054W, YLR298C, YML115C,YMR033WYMR308C,YOR201C, YPL002C, YPL009C, YPR129W
6	90304	nucleic acid metabolic process	0.0294	0.05	YBL014C, YBR198C, YCL054W, YLR298C, YMR033W,YOR201C, YPL002C, YPR129W
7	44237	cellular metabolic process	0.0499	0.06	YBL014C, YBR198C, YCL054W, YLR298C, YML115C, YMR033, YMR308C,YOR201C, YPL002C, YPL009C, YPR129W
8	8152	metabolic process	0.0601	0.07	YBL014C, YBR198C, YCL054W, YLR298C, YML115C,YMR033W, YMR308C,YOR201C, YPL002C, YPL009C, YPR129W
9	6139	nucleobase containing compound metabolic process	0.0644	0.06	YBL014C, YBR198C, YCL054W, YLR298C, YMR033W, YOR201C, YPL002C, YPR129W

Conclusion

In real life situation, there is a need for finding set of genes which are correlated to the given query and not similar to the given query.The ratio based similarity measure defined in

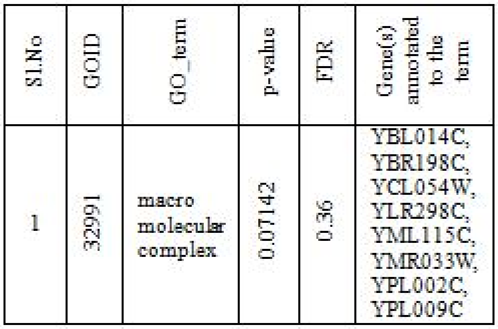
DOI: 01.IJIT.3.1.1015

TABLE V. MOLECULAR FUNCTION OF ADDITIVE BICLUSTER



Sl.No	GOID	GO_term	p-value	FDR	Gene(s) annotated to the term
1	16435	rRNA (guanine) methyltransferase activity	7.06E-05	0	YCL054W, YOR201C
2	8649	rRNA methyltransferase activity	0.00023	0	YCL054W, YOR201C
3	8173	RNA methyltransferase activity	0.01001	0.01	YCL054W, YOR201C
4	16251	general RNA polymerase II transcription factor activity	0.0178	0.01	YBR198C, YMR033W
5	8757	S-adenosylmethio nine-dependent methyltransferase activity	0.05989	0.02	YCL054W, YOR201C
6	8168	methyltransferase activity	0.08763	0.03	YCL054W, YOR201C

TABLE VI. CELLULAR COMPONENT OF ADDITIVE BICLUSTER



Sl.No	GOID	GO_term	p-value	FDR	Gene(s) annotated to the term
1	32991	macro molecular complex	0.07142	0.36	YBL014C, YBR198C, YCL054W, YLR298C, YML115C, YMR033W, YPL002C, YPL009C

SIMBIC+ is efficient in extracting highly correlated biclusters which helps to identify genes with more biological significance. Thus SIMBIC+ query driven biclustering model outperforms MSB and SIMBIC models in terms of biological significance.

REFERENCES

[1] J. Bagyamani, K. Thangavel, “SIMBIC: SIMilarity based BIClustering of Expression Data”. Information Processing and Management Communications in Computer and Information Science, 70, pp. 437-441, 2010.

[2] J. Bagyamani, Thangavel K and Rathipriya R., ‘Biological Significance of Gene Expression Data using Similarity based Biclustering Algorithm’, International Journal of Biometrics and Bioinformatics (IJBB), Vol. 4(6), pp 201-216, 2011.

[3] Y. Cheng, G.M Church, “Biclustering of expression data”. Proceedings of 8th International Conference on Intelligent Systems for Molecular Biology, ISMB-00, pp. 93-103, 2000.

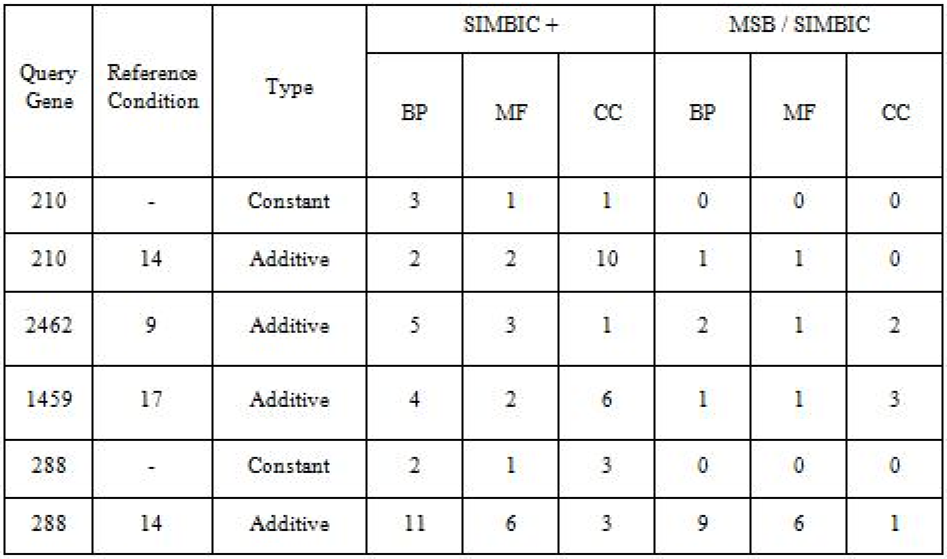
[4] X. Liu and L. Wang, “Computing maximum similarity biclusters of gene expression data”, Bioinformatics, 23(1), pp. 50-56, 2007.

[5] S.C. Madeira and A.L Oliveira. “Biclustering algorithms for biological data analysis: a survey”. IEEE Transactions on Computational Biology and Bioinformatics,1(1), pp. 24-45, 2004.

[6] Roy Varshavsky, Assaf Gottlieb, Michal Linial and David Horn. “Novel Unsupervised Feature Filtering of Biological Data”. Bioinformatics, 22(14), e507-e513, 2006.

[7] A. Tanay, R. Sharan and R. Shamir. “Biclustering Algorithms: A Survey”. Handbook of Computational Molecular Biology, 2004.

TABLE VII. COMPARISON OF GO ENRICHMENT OF BICLUSTERS OF YEAST DATASET OBTAINED BY SIMBIC+ AND MSB



Query Gene	Reference Condition	Type	SIMBIC +			MSB / SIMBIC		
			BP	MF	CC	BP	MF	CC
210	-	Constant	3	1	1	0	0	0
210	14	Additive	2	2	10	1	1	0
2462	9	Additive	5	3	1	2	1	2
1459	17	Additive	4	2	6	1	1	3
288	-	Constant	2	1	3	0	0	0
288	14	Additive	11	6	3	9	6	1